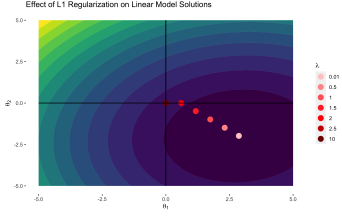
Introduction to Machine Learning

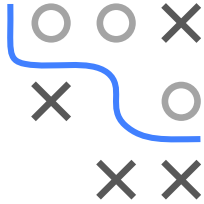
Lasso Regression



Learning goals

- Know lasso regression ($L1$ penalty)
- Know the properties of $L1$ regularization

LASSO REGRESSION



Another shrinkage method is the so-called **lasso regression** (least absolute shrinkage and selection operator), which uses an $L1$ penalty on $\boldsymbol{\theta}$:

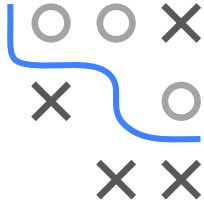
$$\begin{aligned}\hat{\theta}_{\text{lasso}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} |\theta_j| \\ &= \arg\min_{\boldsymbol{\theta}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) + \lambda \|\boldsymbol{\theta}\|_1\end{aligned}$$

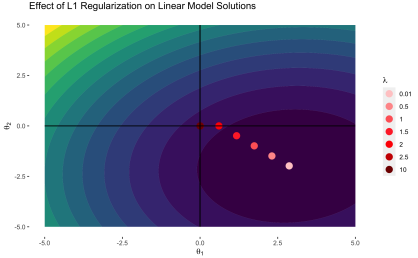
Optimization is much harder now. $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is still convex, but in general there is no analytical solution and it is non-differentiable.

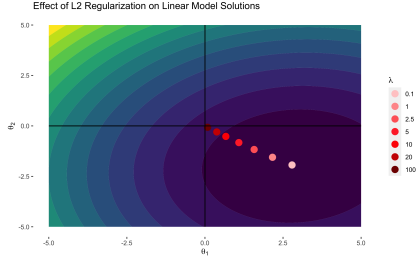
LASSO REGRESSION

Let $y = 3x_1 - 2x_2 + \epsilon$, $\epsilon \sim N(0, 1)$. The true minimizer is $\theta^* = (3, -2)^T$.

Left plot shows effect of $L1$ regularization, right plot shows corresponding with $L2$ for comparison:



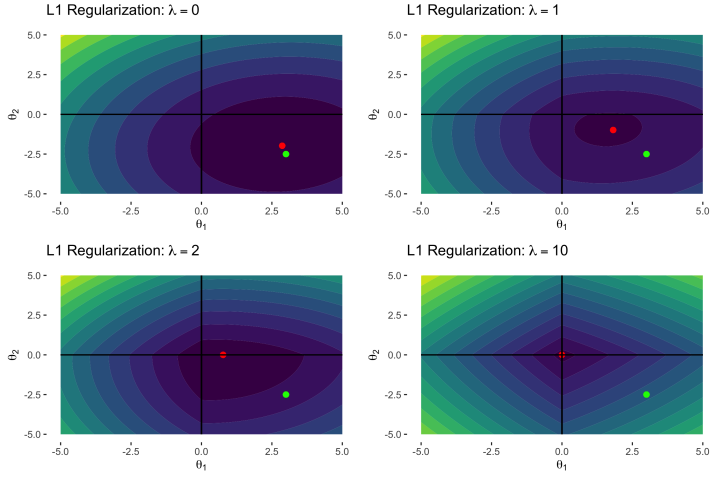




With increasing regularization, $\hat{\theta}_{lasso}$ is pulled back to the origin, but takes a different “route”.

LASSO REGRESSION

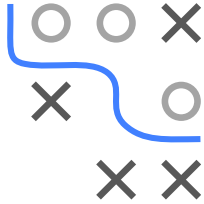
Contours of regularized objective for different λ values.

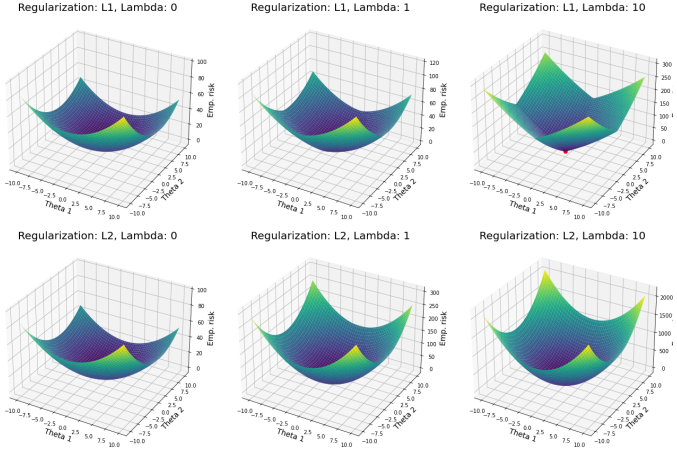


Green = true minimizer of the unreg.objective and red = lasso solution.

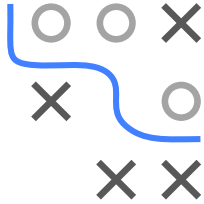
LASSO REGRESSION

Regularized empirical risk $\mathcal{R}_{\text{reg}}(\theta_1, \theta_2)$ using squared loss for $\lambda \uparrow$. $L1$ penalty makes non-smooth kinks at coordinate axes more pronounced, while $L2$ penalty warps $\mathcal{R}_{\text{reg}}$ toward a “basin” (elliptic paraboloid).





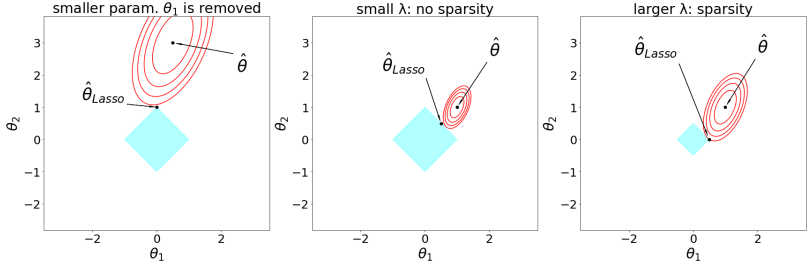
LASSO REGRESSION



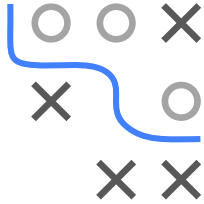
We can also rewrite this as a constrained optimization problem. The penalty results in the constrained region to look like a diamond shape.

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right)^2 \text{ subject to: } \|\boldsymbol{\theta}\|_1 \leq t$$

The kinks in $L1$ enforce sparse solutions because "the loss contours first hit the sharp corners of the constraint" at coordinate axes where (some) entries are zero.



*L*1 AND *L*2 REG. WITH ORTHONORMAL DESIGN

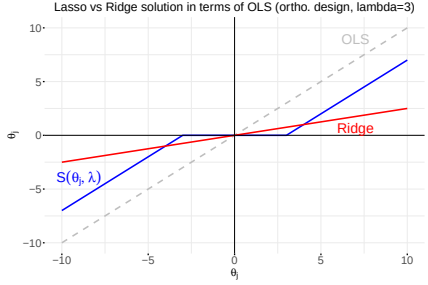


For special case of orthonormal design $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}$ we can derive closed-form a solution in terms of $\hat{\theta}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{y}$:

$$\hat{\theta}_{\text{lasso}} = \text{sign}(\hat{\theta}_{\text{OLS}})(|\hat{\theta}_{\text{OLS}}| - \lambda)_+ \quad \text{(sparsity)}$$

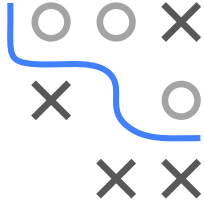
Function $S(\theta, \lambda) := \text{sign}(\theta)(|\theta| - \lambda)_+$ is called **soft thresholding** operator:
For $|\theta| < \lambda$ it returns 0, whereas params $|\theta| > \lambda$ are shrunken toward 0 by λ.
Comparing this to $\hat{\theta}_{\text{Ridge}}$ under orthonormal design:

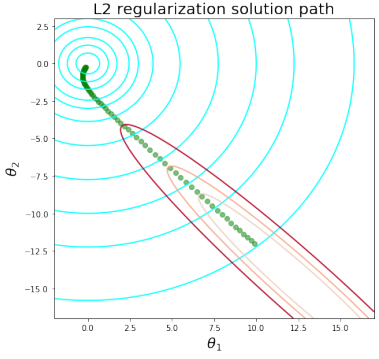
$$\hat{\theta}_{\text{Ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y} = ((1+\lambda)\boldsymbol{I})^{-1}\hat{\theta}_{\text{OLS}} = \frac{\hat{\theta}_{\text{OLS}}}{1+\lambda} \quad \text{(no sparsity)}$$

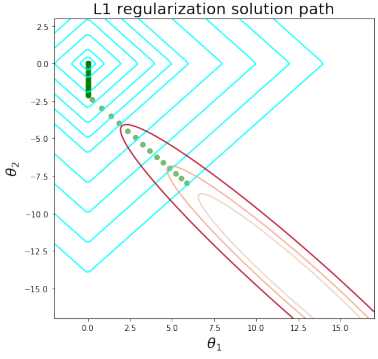


COMPARING SOLUTION PATHS FOR *L1*/*L2*

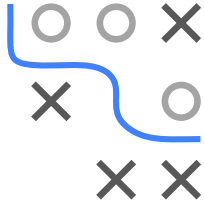
- Ridge regression results in a smooth solution path with non-sparse parameters
- lasso regression induces sparsity, but only for large enough λ







SUPPORT RECOVERY OF LASSO



In which cases can the lasso select the true support of $\boldsymbol{\theta}$? This can be formalized as sign consistency (different from ℓ_2 consistency!):

$$\mathbb{P}\big(\text{sign}(\hat{\theta}) = \text{sign}(\boldsymbol{\theta})\big) \to 1 \text{ as } n \to \infty$$

Suppose the true DGP given a partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\top$ is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \varepsilon \text{ with } \varepsilon \sim (0, \sigma^2\boldsymbol{I})$$

and only $\boldsymbol{\theta}_1$ is non-zero. Let $\mathbf{X}_1$ denote the $n \times q$ matrix with the relevant features and $\mathbf{X}_2$ the matrix of noise features. It can be shown that $\hat{\theta}_{lasso}$ is sign consistent under a **irrepresentable condition** ▸ Zhao and Yu, 2006:

$$|(\mathbf{X}_2^\top \mathbf{X}_1)(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\text{sign}(\boldsymbol{\theta}_1)| < \mathbf{1}$$

In fact, lasso can only be sign consistent if this condition holds.
Intuitively, the irrelevant variables in $\mathbf{X}_2$ must not be too correlated with (or *representable* by) the informative features ▸ Meinshausen and Yu, 2006.